Kísérletek statisztikai és hibrid magyar–angol és angol–magyar fordítórendszerek megvalósítására[1]

Novák Attila, Prószéky Gábor

MorphoLogic
1116 Budapest, Kardhegy u. 5.
{novak,proszeky}@morphologic.hu

Kivonat: Cikkünkben két olyan kísérletről számolunk be, amelyek arra irányultak, hogy a tisztán szabály alapú MetaMorpho rendszerünknél jobb minőségű fordításokat hozzunk létre. Két ilyen rendszer készült: az egyik rendszerben a Moses statisztikai dekódert használtuk a MetaMorpho által előállított fordítások rangsorolására, illetve a részleges fordításokból teljes fordítások összeállítására; a másik kísérleti rendszer egy tisztán statisztikai morfémaalapú magyar–angol fordítórendszer volt. Előbbi rendszerünkkel a tisztán szabály alapú rendszernél kicsit jobb minőségű fordítást kaptunk, az utóbbi azonban gyengébb eredményeket produkált.

1 Bevezetés

A MorphoLogic MetaMorpho fordítórendszere (Novák, Tihanyi & Prószéky, 2008) egy sok emberévtizednyi munkával létrehozott szabályalapú fordítóprogram, amely a magyar és az angol nyelv között mindkét irányban képes fordítani. Időközben létrejöttek ezen nyelvpár tagjai közötti automatikus fordítást kínáló más kísérleti és üzleti alkalmazások, illetve online szolgáltatások is. Ezek között megjelentek a statisztikai gépi fordítási paradigma keretében készült rendszerek is, ám ha az anonimizált gépi fordítások szubjektív emberi minőségi rangsorolását tekintjük mércének, mind a mai napig a MetaMorpho kínálja a legjobb minőségű fordítást. Ebben a cikkben két olyan kísérletről számolunk be, amelyekben a MetaMorphóénál jobb minőségű fordítást produkáló fordítórendszereket próbáltunk létrehozni.

Az eredeti MetaMorpho rendszerben a lehetséges fordítási opciók közötti választás sok esetben nem feltétlenül optimális. Ha a rendszerbe épített mondatelemzőnek sikerült teljes elemzést előállítania a lefordítandó mondathoz, akkor egyszerűen a legelsőként előálló elemzésnek megfelelő fordítást adja vissza, ahelyett hogy esetleg több lehetséges fordítást előállítana, és azok közül választaná ki a legjobbat. Abban az esetben pedig, amikor nem áll elő a fordítandó mondathoz teljes elemzés, és a program részfordításokból próbál a teljes mondatot lefedő fordítást összeállítani, a részfordítások kiválasztásánál nem ellenőrzi, hogy az egyes fordításrészletek a célnyelven

[1] Ehhez a kutatáshoz az Európai Bizottság részleges támogatást nyújtott az EuroMatrix (FP6-IST-5-034291-STP) projektum keretében. Szeretnénk köszönetet mondani Laki Lászlónak és Siklósi Borbálának statisztikai fordítórendszerünk létrehozásában való közreműködésükért.

mennyire jól illeszkednek egymáshoz. Ezért úgy döntöttünk, hogy létrehozunk egy olyan kísérleti hibrid fordítórendszert, amelyben mind a teljes fordítások rangsorolására, mind a részfordítások kiválasztására és azokból a teljes fordítás összeállítására a MetaMorpho eredeti algoritmusa helyett a Moses statisztikai dekódert használjuk (Koehn és munkatársai, 2007).

Létrehoztunk emellett egy teljesen statisztikai alapon működő alternatív fordítórendszert is (szintén a Moses felhasználásával), amelyben a hagyományos szóalapú megoldás helyett morfématokeneket használtunk. Ezt a megoldást a magyar és az angol nyelv közötti alapvető strukturális különbségek és az ezek által okozott szómegfeleltetési (alignment) problémák motiválták, amelyek a jelenleg elterjedt frázis alapú statisztikai gépi fordítási paradigmában alapvetően behatárolják az angol–magyar viszonylatban elérhető fordítási minőséget. Sajnos azonban utóbbi rendszerünk nem bizonyult sikeresnek: az általa generált fordítások minősége mind a BLEU-pontszám, mind a szubjektív emberi megítélés szempontjából messze elmaradt a szabályalapú rendszer (és a szóalapú statisztikai rendszerek) fordításainak minőségétől.

2 A MetaMorpho fordítórendszer

A MorphoLogic MetaMorpho szabályalapú fordítórendszere strukturálisan különbözik a legelterjedtebb szabályalapú fordítórendszerektől: nem tartalmaz külön transzfer komponenst. Nyelvtana (beleértve a lexikont is) olyan mintapárokból áll, amelyeknek egyik tagját a forrásmondat (alulról felfelé történő) elemzésekor használja a fordítórendszer mondatelemzője, és az ehhez tartozó célnyelvi mintát (vagy több célnyelvi minta valamelyikét) felhasználva generálja az adott forrásnyelvi mondatrészlet célnyelvi megfelelőjét a fordítás (felülről lefelé történő) generálásakor. A mintapárok tagjai jegyekkel kibővített kontextusfüggő szabályok. A nyelvtan architektúrája teljesen homogén: az általános szerkezeti szabályoktól a többé-kevésbé idiomatikus frázisokon keresztül a teljesen lexikalizált szótári tételekig minden nyelvi elemet és azok fordítását azonos módon ábrázolja, ezek csak az egyes elemek alulspecifikáltságának mértékében különböznek egymástól.

A célnyelvi szerkezetek létrehozása és a lexikai elemek beillesztése nem igényel utólagos transzfer műveletet: a forrásnyelvi elemzési fa részstruktúráinak az alkalmazott szabálypárok szerint megfelelő célnyelvi struktúrákat csak ki kell olvasni, és azokat a célnyelvi szóalak-generátor közvetlenül fordítássá alakítja.

A MetaMorphóban a forrásnyelvi szöveg elemzése az alábbi lépésekből áll. Az első lépés a szöveg mondatokra bontása. Ezt a szavakra bontás, azaz tokenizálás és a tokenek morfológiai elemzése követi, amely morfoszintaktikai jegyvektorokat rendel a tokenekhez. Ezután következik a többértelmű tokensorozatok által alkotott háló elemzése a nyelvtan forrásoldali szabályainak felhasználásával. A nyelvtanban jegyeket használunk egyrészt az elemzett szövegre vonatkozó lexikai, morfoszintaktikai és vonzatkeret-információk tárolására, másrészt arra, hogy az elemzési, illetve generáló szabályok alkalmazhatóságára vonatkozó megszorításokat tegyünk (pl. másként fordítunk egy igét, ha az alanya élő, mint ha nem az).

Amikor az elemzés kész, és nem marad több illeszthető elemzési szabály, a fordítás a forrásnyelvi mondat elemzési fáját felülről (a mondatszimbólumtól kezdve) bejárva az egyes forrásnyelvi részstruktúráknak megfelelő célnyelvi struktúrák kombinálásával, a bennük szereplő lexikai és morfoszintaktikai jegyegyüttesek interpretációjával áll elő. A forrásnyelvi szabályok bármelyikéhez egynél több célnyelvi szabály is tartozhat. Az adott esetben alkalmazandó célnyelvi megfelelő kiválasztásakor a rendszer az adott forrásnyelvi szabály alkalmazásakor kitöltött jegyekre tett megszorításokra támaszkodik.

A klasszikus transzfer alapú fordítóktól eltérően, a MetaMorphóban a fordításkor alkalmazandó szórendi átrendezéseket is a forrásnyelvi szöveg elemzése során alkalmazott szabályok és az elemzési fában kitöltött jegyek tulajdonképpen már elemzési időben meghatározzák. A kimenet generálásakor csak a már meghatározott és átrendezett struktúrák szöveggé alakítása történik. A generált célnyelvi fa terminális pontjain levő morfoszintaktikai és lexikai jegyegyüttesek interpretálását a célnyelvi szóalak-generátor végzi, amely a megfelelő célnyelvi szóalakokat előállítja.

A többértelműségek kezelése a tisztán szabályalapú rendszerekben mindig nehéz. A MetaMorpho két módszert alkalmaz a nem kívánt többértelműségek kiszűrésére: vagy magas szintű heurisztikákat használ az alternatívák közötti választásra (pl. egy összetevőnek vonzatként való elemzését preferálja a szabad határozóként való elemzés helyett), vagy a specifikusabb szabályok explicit módon blokkolják az adott esetben nem alkalmazandó általánosabb szabályok alkalmazását.

Általában a MetaMorpho csak az első sikeres elemzéshez tartozó első lehetséges fordítást állítja elő. Kellően hosszú, és megfelelő számú lehetséges strukturális többértelműséget tartalmazó fordítandó mondatok esetében azonban így is előfordulhat, hogy elemzés közben túl sok hipotézis áll elő. Eredetileg erre a problémára az volt a megoldás, hogy az elemző egyszerűen leállt azon a ponton, amikor egy beállított időkorlátot túllépve túl sok időt töltött egy mondat elemzésével. Ez a megoldás ugyan biztosítja azt, hogy a fordítórendszer válaszideje minden bemenetre korlátos maradjon, azonban ennek a megoldásnak az eredményeképpen az ilyen, túl hosszú mondatokra olyan fordítás jött létre, amely a mondat végén lefordítatlanul maradt szavakat tartalmazott. Erre a problémára jobb megoldást sikerült találni azzal, hogy a túl hosszúnak tűnő mondatokat már a mondatokra bontás során rövidebb egységekre bontjuk (a korábbinál agresszívebb módon), és így már szinte egyáltalán nem fordul elő, hogy szükség lenne az elemzés idő előtti megszakítására, és ennek megfelelően sokkal ritkábban maradnak lefordítatlan részek a fordításban.

3 A hibrid fordítórendszer

Elemzés közben a MetaMorpho mondatelemzője hierarchikusan egymásba épülő részleges szintaktikai struktúrákat állít elő. Ha nem sikerül teljes elemzést találni az adott lefordítandó mondathoz, akkor a MetaMorpho jobb híján egy olyan heurisztikát alkalmaz, amely ezekből a részleges struktúrákból egy a teljes bemenő mondatot mintegy mozaikszerűen lefedő sorozatot kiválasztva állítja elő a fordítást. Az így előálló fordítások általában nem optimálisak, mert a teljes elemzés hiányában bizonyos strukturális (pl. az egyeztetéssel kapcsolatos) információ elvész.

3.1 A névmástörlés

A magyar–angol fordítási irányban a magyar névmások kiesése (az ún. pro-drop) további problémát jelent, amikor részfordításokból próbáljuk a teljes fordítást összerakni. Mivel az alany számát és személyét, vagy tárgyas igék esetében a tárgy határozottságát az igeragok általában önmagukban pontosan jelzik. Az explicit alanyi és tárgyi névmások tehát a magyarban általában elhagyhatók, és gyakran el is hagyjuk őket (hacsak nem állnak fókuszban, vagy egyéb módon kiemelten hangsúlyosak). A probléma az, hogy pontosan ugyanazokat az igealakokat használjuk kitett teljes alany és tárgy mellett, mint amiket az elhagyott névmások esetében. Ebben az esetben azonban ugyanezek az igei végződések nem tartalmaznak inkorporált névmást, és hiba, ha a fordításban névmás jelenik meg.

Hallja.	*He/she/it hears him/her/it.*
Fred hallja a doktort.	*Fred hears the doctor.*

Puszta (egyszavas) magyar igealakok fordításaként a MetaMorpho kizárólag olyan angol frázisokat generál, amelyek explicit alanyi névmást tartalmaznak (illetve határozott tárgyas igealakok, pl. a *hallja* esetében tárgyi névmást is: *he hears it*), mert az igéket a nyelvtanban kizárólag a vonzataikat is tartalmazó lexikai minták reprezentálják. Ennek következtében fölösleges beszúrt névmások jelenhetnek meg azokban a mozaikszerűen összerakott fordításokban, ahol testes alany, illetve tárgy is szerepel a mondatban, abban az esetben, ha az algoritmus olyan forrásnyelvi részmondat fordítását is felhasználja, amelyben explicit alany vagy tárgy nem szerepelt.

Hasonló jelenség figyelhető meg a harmadik személyű birtokos szerkezetek esetében (itt birtokos névmások jelenhetnek meg birtokos szerkezetek helyett):

háza	*his house*
Fred háza.	*Fred's house.*

Egy példa a fentiekre a következő fordítás:

Bemenet: *A repülő objektumok + nem viselkednek teljes mértékben úgy, mint ahogy az az ősi gravitációs törvény + alapján + elvárható + lenne.*

MMO: *The flying objects + they do not behave in a full measure the way that ancient gravitational law + his basis + can be expected + he would be.*

3.2 A Moses dekóder bevetése

Az eredeti részfordítás-kombináló algoritmus nem használ célnyelvi nyelvmodellt arra, hogy a lehetséges részekből összerakott fordításokat rangsorolja. Kísérleteinkben az eredeti algoritmust statisztikai modellel helyettesítettük. A hibrid rendszerben a MetaMorphót a nyílt forráskódú Moses statisztikai dekóderrel kombináltuk (Koehn és munkatársai, 2007): a szabályalapú komponens által előállított részfordításokat, illetve teljes fordításokat tartalmazó frázistáblából a Moses dekóder állít össze és

választ célnyelvi nyelvmodell felhasználásával optimalizált fordítást. Azt reméltük, hogy így az eredetinél jobb minőségű fordítást kapunk ezekben az esetekben. A MetaMorpho elemzőjét kiegészítettük egy olyan felülettel, amely az elemzés közben létrejött összes részstruktúrát a lehetséges fordításaival együtt ki tudja írni a Moses dekóder frázistáblájának megfelelő formátumban.

Ennek felhasználásával aztán a Moses dekóder segítségével generáltunk célnyelvi nyelvmodell felhasználásával optimalizált fordítást az eredeti fordítandó mondatokra. Mivel jobb becslésünk nem volt a fordítási valószínűségekre, egyenletes eloszlást feltételeztünk a frázistáblában az egyes frázisok lehetséges alternatív fordításai felett, és a Moses konfigurációjában zérus súlyt rendeltünk a fordítási modellhez. Lexikalizált torzítási modellt sem használtunk (a statisztikai fordítási zsargonban a szórendi átrendezést nevezik torzításnak). Így a dekóder a célnyelvi nyelvmodell által a fordításhoz rendelt pontszám alapján rangsorolja a fordításokat. Kísérleteinkben 5-gram (5 szavas) nyelvmodellt használtunk, amelyet a Hunglish korpusz (Halácsy és munkatársai, 2005) jogi és irodalmi részéből generáltunk. Sajnos nagyobb egynyelvű korpuszokból generált nyelvmodellek előállítását és betöltését a használt tesztgépben levő operatív memória menyisége nem tette lehetővé. [2]

Számos paraméterbeállítással és frázistábla-építési módszerrel kísérleteztünk. A teljes elemzéssel rendelkező mondatok esetében a részfordítások felvétele a frázistáblába a fordítási minőség egyértelmű romlásához vezetett. Ugyanakkor – nem meglepő módon – az összes lehetséges teljes fordítás felvétele a frázistáblába (ha volt a mondatnak sikeres teljes elemzése) és a legjobb fordítás nyelvmodell segítségével való kiválasztása a MetaMorpho-alapértelmezéssel, azaz az első sikeres elemzésnek megfelelő fordítást kiíró megoldással szemben egyértelműen javította a fordítás minőségét. A dekóder konfigurációs fájljában meg kellett növelnünk a maximális megengedett frázisméret értékét az alapbeállításról ahhoz, hogy a dekóder a teljes mondatfordításokat is használja (ellenkező esetben nagyon drasztikusan romlott a fordítások minősége).

Szintén kedvező hatása volt, ha azokhoz a frázisokhoz, amelyeknek a fordítása esetleg felesleges névmást tartalmazott, olyan alternatív fordításokat is generáltunk a frázistáblába, amelyekben a névmások nem szerepeltek, mert ez tényleg csökkentette a fordító által generált felesleges névmások számát.

Míg a MetaMorpho eredeti részfordítás-összerakó algoritmusa soha nem próbálja meg átrendezni a generált darabokat, a hibrid rendszerben kísérleteztünk különböző torzítási (pontosabban: szórend-átrendezési) beállításokkal, hiszen ez a lehetőség benne van a Mosesben. (Azért nem egészen „ingyenes” ez a szolgáltatás: az átrendezés megengedése drasztikusan növeli a dekódoláshoz – az optimális fordítás kiválasztásához – szükséges időt.) Azt találtuk, hogy ha nem adtunk büntetőpontokat a szórendi átrendezésekért a dekódernek, akkor határozottabban rosszabb minőségű fordításokat kaptunk. Az alapbeállításban szereplő torzítási büntetés (a torzítási büntetést és a nyelvmodell által adott pontszámot azonos súllyal vettük figyelembe), és megengedett maximális mozgatás (d=6, azaz 6 szón átívelő mozgatás megengedése) gyak-

[2] Lehetséges megoldások erre a problémára (amellett, hogy több memóriát teszünk a gépbe): alacsonyabb rendű nyelvmodell használata (ezzel persze a távolabbi függőségek ellenőrzését csökkentjük), az egyszeri előfordulások elhagyása és/vagy a nyelvmodell szótárának a leggyakoribb frázisokra korlátozása.

ran olyan fordításokat eredményezett, amelyekben a fordításként generált mondat végén teljesen elkeveredett fordításdarabok sorakoztak. A legjobb eredményt – a BLEU-pontszám tekintetében is – abban az összeállításban kaptuk, amelyikben az átrendezést teljesen megtiltottuk (d=0), annak ellenére, hogy ez sokszor szórendileg szerencsétlenebb fordítást eredményezett, különösen a magyar–angol fordítási irányban, ha a fordítandó magyar mondatnak a végén állt az ige. Az átrendezés letiltása a dekódolási időt is tizedére csökkentette.

Az alábbi mondat esetében látható egyrészt a feleslegesen generált névmások elhagyásának kedvező hatása, másrészt itt a hibrid fordító egyébként is sokkal érthetőbb fordítást generált, annak ellenére, hogy az egyik ige nem a megfelelő helyre került a fordításban:

Bemenet: *„Az anomáliáért a sötét anyag lehet felelős, amely talán akár egészen a Föld közelében is megtalálható”, írja Adler.*

MMO: *The dark substance, which the Earth is entirely in his neighbourhood even possibly, may be responsible for the anomaly can be found, Adler writes it.*

MMO+Moses: *The dark substance may be responsible for the anomaly, that possibly even all near the Earth can be found, Adler writes.*

3.3 Eredmények

A kísérleti összeállításokat a 2009-es athéni EACL konferencia mellett rendezett *Fourth Workshop on Statistical Machine Translation*-re kiadott angol–magyar tesztkészleten teszteltük (Callison-Burch és munkatársai, 2009).

Legeredményesebbnek a következő kísérleti összeállítás bizonyult:

- a frázistáblát kiegészítettük olyan alternatív részfordításokkal is, amelyekből töröltük a beszúrt névmásokat,
- a Moses dekódert úgy paramétereztük, hogy ne rendezze át az összetevők sorrendjét,
- azokra a mondatokra, amelyekre a MetaMorpho teljes fordítást adott, nem használtuk a részfordításokat, hanem pusztán a teljes fordítások rangsorolására használtuk a dekódert.

Az utóbbi összeállítással mindkét fordítási irányban a puszta MetaMorphónál kissé jobb minőségű fordításokat sikerült elérni mind a BLEU-pontszám, mind a szubjektív emberi megítélés szempontjából, azonban a javulás mértéke elmaradt a várakozásainktól (BLEU: magyar–angol irányban 9,96 → 10,10; angol–magyar irányban 8,13 → 8,44). Az alábbi táblázatban összefoglaltuk az eredeti MetaMorpho rendszer és néhány hibrid összeállítás által generált fordítások BLEU-pontszámait:

1. táblázat: A fordítások és azok BLEU-pontszámai.

magyar–angol	
MetaMorpho	9.96
d=6, nincs átrendezési büntetés, teljes elemzésnél is lehet összerakás	9.62
d=6, van átrendezési büntetés, teljes elemzésnél nincs összerakás	9.70
d=0, nincs átrendezés, teljes elemzésnél nincs összerakás, névmástörlés	10.10
angol–magyar	
MetaMorpho	8.13
d=6, van átrendezési büntetés, teljes elemzésnél nincs összerakás	8.22
d=0, nincs átrendezés, teljes elemzésnél nincs összerakás	8.44

4 Morfémaalapú statisztikai fordítórendszer

A magyar–angol fordítási irányban kísérleteztünk egy további fordítórendszerrel is, amelyben a szabályalapú komponenst mellőzve, a statisztikai nyelvmodelleket algoritmikus morfológiai elemzővel és szófaji egyértelműsítővel előállított morfémaalapú reprezentáció felhasználásával állítottuk elő. Ebben a rendszerben szintén a Moses dekódert használtuk.

4.1 A rendszer felépítése

A tanítókorpusz magyar oldalát a *Humor* morfológiai elemzővel (Prószéky & Novák, 2005) elemeztük és tövesítettük, és a *Hunpos* szófaji egyértelműsítővel (Halácsy, Kornai & Oravecz, 2007) egyértelműsítettük. Az angol oldal egyértelműsítésére a *CRFTagger*-t (Phan, 2006) használtuk, és a *morpha* elemzővel tövesítettünk (Minnen, Carroll & Pearce, 2001). Az utóbbinak megfelelő *morphg* morfológiai generátorral állítottuk elő célnyelvi fordítások felszíni alakjait. Sajnos a *morpha* elemző nem különbözteti meg a létige nem harmadik személyű alakjait a harmadik személyűektől, ezért ezt a hibát javítanunk kellett ahhoz, hogy a kimeneten a létige helyes alakja generálódjon.

Rendszerünkben a tokenek nem szavak, hanem morfémák voltak. Az alábbi példa a tanítókorpusz egy mondatát mutatja a rendszerben használt morfémaalapú reprezentációban.

Magyar: *a[det] 137[szn] apró[mn] csillag[fn] [ela] álló[mn] spirál[fn] meg+[ik] duplázódik[ige] [me3] .[punct]*

Angol: *the_dt spiral_nn of_in 137_cd tiny_jj star_nn s_nns double_vb ed_vbd itself_prp ._.*

Megközelítésünket több tényező motiválta. Egyrészt a magyarban a szavaknak több ezer lehetséges toldalékolt alakja van, és nincs az a korpusz, amelyben példaként

szerepelne minden szó minden lehetséges alakja (vagy akár csak a leggyakoribbak). Ezért az adatorientált megközelítés lépten-nyomon abba a problémába ütközik, hogy hiányzik az éppen szükséges adat, ha a tokenek szóalakok. Másrészt rendszeresen kötött morfémák felelnek meg a magyarban angol funkciószavaknak (pl. elöljárószóknak, birtokos és egyéb névmásoknak). Emellett rendszeres morfémasorrendi különbségek is vannak: az angol prepozícióknak a magyarban megfelelő toldalékok, illetve névutók követik, és nem megelőzik a névszói csoportokat, ugyanez igaz a birtokos névmásokra (és a nekik megfelelő birtokos ragokra), illetve az alanyi névmásokra (amelyeknek a magyarban leggyakrabban csak az igei személyragok felelnek meg).

Ezek a tényezők elég súlyos problémákat okoznak már a statisztikai fordító betanításához használt tanítókorpuszban az egymásnak megfeleltethető szóalakok összepárosítását végző Giza++ számára is, illetve jelentősen csökkentik a szóalapú fordítórendszer általánosítóképességét. Azt reméltük, hogy morfémaalapú rendszerünk frappánsan megoldja ezeket a problémákat.

A frázistáblát az alapértelmezett *grow-diag-final* heurisztikával állítottuk elő a Giza++ szóösszerendelésekből, amelyet a tanítókorpusz morfémaalapú reprezentációjából állítottunk elő. Ebben a rendszerben használtunk lexikalizált átrendezési táblát, a torzítási paramétert az alapbeállításon hagytuk. A rendszerben 5-gramos célnyelvi nyelvmodellt használtunk (ebben az esetben ez öt morfémát, nem öt szót jelent). Sajnos ebben az esetben is csak korlátozott méretű korpuszból tudtunk nyelvmodellt építeni a tesztrendszer korlátozott memóriakapacitása miatt. A rendszer betanításához a *Hunglish* korpusz irodalmi és jogi részét használtuk, a tesztkorpusz azonos volt a hibrid rendszer esetében használttal.

A MERT paraméteroptimalizációs eljárást (Och, 2003) úgy futtattuk, hogy az a korpuszból kiválasztott hangolókészleten kapott morfémaalapú BLEU-pontszámot próbálta optimalizálni. Az optimalizálás több napig futott.

4.2 Eredmények

A rendszer tesztelésekor először a morfémaalapú BLEU-pontszámot optimalizáló MERT eljárás által javasolt paraméterbeállításokat használtuk. A célnyelvi angol szóalakokat a *morphg*-vel állítottuk elő a dekóder által előállított morfémaalapú fordításokból. Számítottunk rá, hogy a morfémaalapú rendszer új problémával szembesít majd minket: olyan helyekre fognak keveredni morfémák, ahol normális esetben nem fordulhatnának elő, és így nem tudunk majd értelmes szóalakot generálni az adott morfémasorozatból. Így is lett. Ezekben az esetekben egyszerűen kihagytuk a rossz helyre került morfémát, bár ez nyilván nem optimális megoldás.

Sajnos ez az összeállítás várakozásainkkal ellentétben nem produkált nagyon biztató eredményeket. A fenti összeállítás a detokenizált kimenetre 7,82-es BLEU-pontszámot adott. Mikor a dekódert újrafuttattuk egy korábbi félbeszakadt MERT folyamat során kapott paraméterekkel, kicsit jobb BLEU-pontszámot kaptunk: 7,95-öt. Ez is sokkal gyengébb volt, mint a MetaMorphóé, de a fordítás emberi megítélése szempontjából még ennél is jelentősebb mértékben elmaradt a minősége a szabályalapú fordítóétól. Nagyjából ugyanez mondható el a rendszer kimenetét szóalapú statisztikai rendszerek által magyar–angol irányban produkált fordításokkal összevet-

ve is: a BLEU-pontszámok különbsége ebben az esetben még nagyobb, és a szubjektív minőség is jelentősen rosszabb a szóalapú rendszerekhez viszonyítva is.

A Giza++ szóösszerendeléseket átnézve azt tapasztaltuk, hogy várakozásainkkal ellentétben a tanítókorpusz morfémákra bontása önmagában nem oldotta meg a szóösszerendelések minőségével kapcsolatos problémákat: az összerendelések még roszszabbak voltak, mint amiket a korpusz minden morfológiai feldolgozás nélküli változatára kaptunk. Ugyanakkor a morfémaalapú megközelítés mindazon hátrányai, amikre előre számítottunk: a nyelvmodellekben és a frázistáblában megragadott lokális függőségek csökkent távolsága annak következtében, hogy a bemenet ugyanakkora szakaszát több token fedi le, mint a szóalapú változatban; a rossz helyre keveredett morfémák stb. valóban bekövetkeztek.

5 Összefoglalás

Cikkünkben a magyar és angol nyelvpár tagjai közt fordító hibrid és morfémaalapú statisztikai kísérleti fordítórendszereinket mutattuk be. Sajnos átütő eredményekről nem számolhattunk be. Ugyan hibrid rendszerünk egyértelműen jobb fordításokat hozott létre, mint a tisztán szabályalapú MetaMorpho rendszer, a javulás mértéke elmaradt várakozásainktól. Morfémaalapú statisztikai fordítórendszerünk pedig egyértelműen nem váltotta be a hozzá fűzött reményeket.

Hivatkozások

1. Callison-Burch, Chris; Philipp Koehn, Christof Monz, Josh Schroeder: Findings of the 2009 Workshop on Statistical Machine Translation In: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Athens, Greece (2009) 1–28
2. Halácsy, Péter; András Kornai, Csaba Oravecz: HunPos – an open source trigram tagger In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Association for Computational Linguistics, Prague, Czech Republic (2007) 209–212
3. Halácsy Péter, Kornai András, Németh László, Sass Bálint, Varga Dániel, Váradi Tamás, Vonyó Attila: A Hunglish korpusz és szótár. In: Csendes D., Alexin Z. (szerk.) Magyar Számítógépes Nyelvészeti Konferencia 2005, Szeged: Szegedi Tudományegyetem, Informatikai Tanszékcsoport. (2005) 134–142
4. Koehn, Philipp; Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst: Moses: Open Source Toolkit for Statistical Machine Translation In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, Association for Computational Linguistics, Prague, Czech Republic, (2007) 177–180
5. Minnen, Guido; John Carroll, Darren Pearce: Applied Morphological Processing of English, Natural Language Engineering, 7(3). (2001) 207–223

6. Novák, Attila; László Tihanyi, Gábor Prószéky: The MetaMorpho translation system. In: Proceedings of the Third Workshop on Statistical Machine Translation at ACL 2008, Columbus, Ohio, (2008) 111–114
7. Och, Franz Josef: Minimum Error Rate Training for Statistical Machine Translation. In: Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL), Sapporo, (2003) 160-167
8. Phan, Xuan-Hieu: CRFTagger: CRF English POS Tagger. (2006) http://crftagger.sourceforge.net/
9. Prószéky, Gábor and Attila Novák: Computational Morphologies for Small Uralic Languages. In: A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund, A. Yli-Jyrä (eds.): Inquiries into Words, Constraints and Contexts Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday, Gummerus Printing, Saarijärvi/CSLI Publications, Stanford. (2005) 116-125